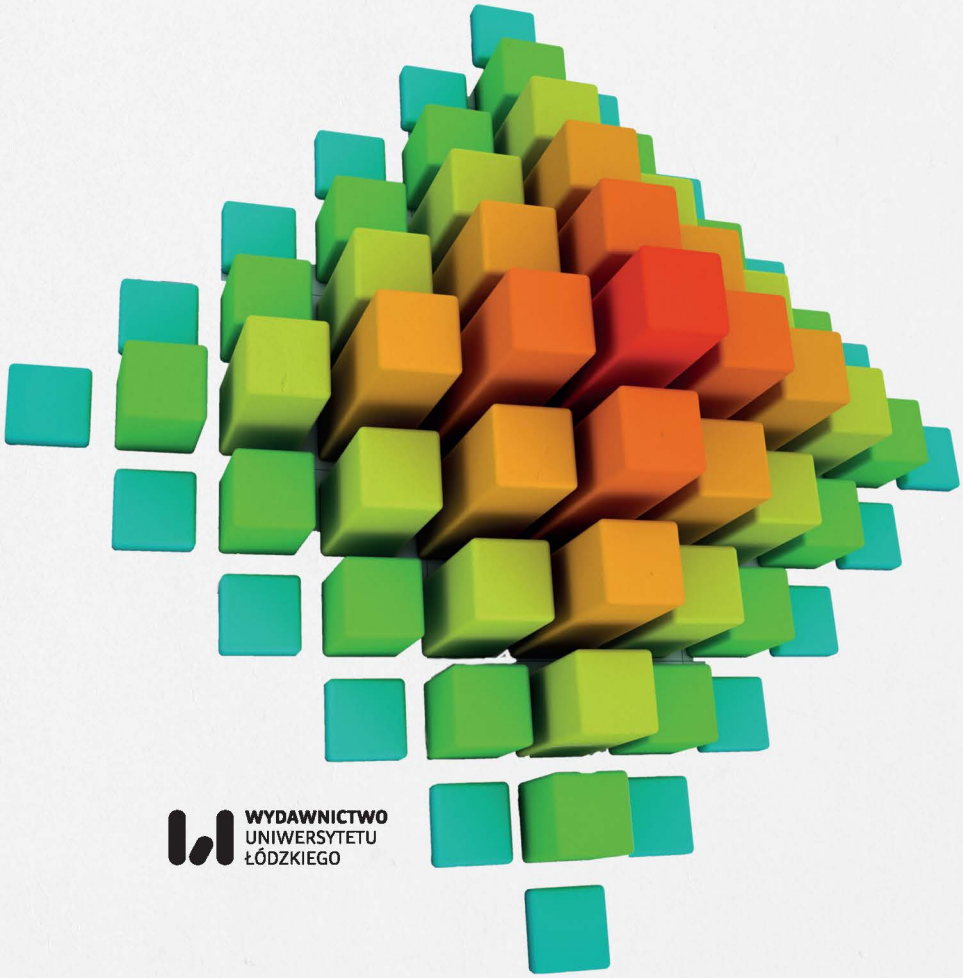
Wiesław Szymczak

Praktyka wnioskowania statystycznego



WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Praktyka wnioskowania statystycznego

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Wiesław Szymczak

Praktyka wnioskowania statystycznego

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Łódź 2018

Wiesław Szymczak – Uniwersytet Łódzki, Wydział Nauk o Wychowaniu
Instytut Psychologii, 91-433 Łódź, ul. Smugowa 10/12

RECENZENT
Grażyna Wieczorkowska-Wierzbińska

REDAKTOR INICJUJĄCY
Urszula Dzieciątkowska

REDAKTOR WYDAWNICTWA UŁ
Katarzyna Gorzkowska

SKŁAD I ŁAMANIE
AGENT PR

PROJEKT OKŁADKI
Katarzyna Turkowska

Zdjęcie wykorzystane na okładce: © Depositphotos.com/benjaminet

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.08579.18.0.M

Ark. wyd. 9,4; ark. druk. 13,375

ISBN 978-83-8142-211-6
e-ISBN 978-83-8142-212-3

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63

Spis treści

Przedmowa

Książkę tę napisałem w celu przypomnienia użytkownikom metod statystycznych ograniczeń, jakie one mają. Są to ograniczenia wynikające bezpośrednio z aksjomatyki, np. teorii testowania hipotez statystycznych, a także ograniczenia będące konsekwencją niedoskonałości używanego oprogramowania statystycznego. Podobnie jak nie istnieje prawdziwy model, tak też nie istnieje program komputerowy całkowicie wolny od błędów. Jednak wskazywanie błędów w oprogramowaniu nie było moim celem.

Publikacja ta nie jest podręcznikiem metod statystycznych. Ma ona na celu zwrócenie uwagi badaczom opracowującym wyniki swoich badań m.in. metodami statystycznymi, że u podstaw każdej stosowanej metody statystycznej leżą założenia, które umożliwiają sformułowanie i udowodnienie pewnych twierdzeń. Twierdzenia te z kolei pozwalają określić właściwości, np. testów statystycznych. Jeśli założenia twierdzenia nie będą spełnione, to samo twierdzenie przestaje być prawdziwe. Konsekwencją rozbieżności między teorią a praktyką statystyki są próby określenia sposobów umożliwiających stosowanie metod statystycznych mimo niespełniania przez materiał empiryczny założeń teoretycznych (np. odporność metod, mierniki oceny wielkości efektu i mnóstwo innych – mniej lub bardziej udanych – pomysłów).

Innym poważnym zagrożeniem poprawności stosowanych metod statystycznych jest dostępność oprogramowania statystycznego. Oczywiste jest, że komputer to tylko maszyna, która policzy prawie wszystko, ale tylko policzy. Używanie zaawansowanego oprogramowania przez osoby nieznające podstaw statystyki często będzie prowadziło do podejmowania irracjonalnych decyzji. Dlatego też zwracam w tej książce uwagę na tzw. kult istotności statystycznej, któremu pod żadnym pozorem nie wolno nam ulegać. Jest on bardzo wygodny, ponieważ – mówiąc brutalnie – zwalnia badacza z myślenia. A przecież to ostatnia rzecz, z której powinniśmy się zwalniać.

Przedstawione w książce zagadnienia nie wyczerpują wszystkich problemów związanych ze stosowaniem metod statystycznych. Należą do nich m.in. błędy w nazewnictwie wynikające z niepoprawnego tłumaczenia terminologii statystycznej czy błędne nazwy pewnych wyników. Sporym problemem są też

różnice w zaimplementowanych szczegółowych rozwiązaniach niektórych metod w różnych programach statystycznych.

Będę szczęśliwy, jeśli choć kilku osobom ułatwię bardziej świadome korzystanie z oprogramowania i metod statystycznych.

Autor

Wstęp

Jedno ze znaczeń słowa „statystyka" brzmi: „nazwa dyscypliny naukowej, będącej gałęzią matematyki i posiadającej własny zestaw narzędzi i metod". Chodzi tu o statystykę matematyczną. Będę ją nazywał statystyką teoretyczną albo teorią statystyki, dla podkreślenia jej dedukcyjnego charakteru. Natomiast stosowanie, wykorzystywanie metod statystycznych w praktyce – to proces indukcyjny. I ta dwoistość – dedukcja w teorii i indukcja w wykorzystaniu – prowadzi do bardzo poważnych komplikacji, które będę chciał zasygnalizować, a niektóre, być może, próbować wyjaśnić.

Jeszcze zdanie usprawiedliwienia dla tytułu książki. Zdecydowałem się na „praktykę wnioskowania statystycznego", gdyż główny akcent położyłem na opis istniejącego stanu stosowania metod statystycznych podczas opracowywania wyników badań ilościowych, głównie w naukach społecznych. Statystyczna analiza danych stosunkowo często wykonywana jest – z punktu widzenia teorii statystyki – nieprawidłowo. Nieprawidłowo w tym sensie, że badacz rzadko sprawdza założenia teoretyczne leżące u podstaw konkretnej metody. Ponadto, ponieważ teoria testowania hipotez statystycznych nie jest pozbawiona wad, powstają różne dziwne protezy mające eliminować te wady, lecz z kolei nie mają one podstaw teoretycznych, co często czyni ich stosowanie „działaniem magicznym".

> Opinie statystyków o statystyce są zwykle skrajnie optymistyczne, a nadużywanie statystyki stanowi *signum temporis* ubiegłych dekad. Jednak euforia wydaje się powoli przemijać i – nie tylko w statystyce – przychodzą chyba czasy większego poczucia rzeczywistości i odpowiedzialności zarazem. Statystyka powinna wracać pomału do punktu wyjścia: formalizowania i analizy wyników badań zjawisk empirycznych. Wymaga to przewartościowania wielu tradycji i odrzucenia wielu mitów.
>
> […]
>
> Mamy nadzieję, że przyczynimy się do tego naszą książką. Chcemy przedstawić ostry konflikt między tym, co możliwe a tym, co potrzebne. Chcemy też przedstawić nikłość powiązań między teorią a praktyką wnioskowania statystycznego, które nieraz ograniczają się do wzajemnych inspiracji.

> […] związki między problemem praktycznym a jego formalnym przedstawieniem są na ogół słabe, a rozwiązanie formalnego problemu może nie mieć racjonalnej interpretacji praktycznej.
>
> […] dochodzenie ojcostwa jest formalnie szczególnym problemem dyskryminacji, ale w praktycznym rozwiązywaniu tego problemu prawie nie korzysta się z teorii.

Powyższe cytaty pochodzą z przedmowy do książki pod redakcją Bromka i Pleszczyńskiej (1988) i przedstawiają realistyczną diagnozę problemów pojawiających się podczas stosowania metod statystyki matematycznej. Jednak diagnoza, iż „przychodzą chyba czasy większego poczucia rzeczywistości […]" wydaje mi się zbyt optymistyczna, Sądzę, że od czasu wydania książki Bromka i Pleszczynskiej bardzo mało zmieniło się w odbiorze i wykorzystaniu statystyki w naukach społecznych. W dalszej części mojej książki wielokrotnie będę wracał do tych problemów, aby skłonić Czytelnika do możliwie precyzyjnych przemyśleń – bardziej w kategoriach merytorycznych niż statystycznych – badanego i rozwiązywanego zagadnienia.

Należy nieco doprecyzować pierwsze zdanie z powyższych cytatów. Otóż, musimy cały czas rozróżniać statystyków teoretyków i statystyków praktyków, a także statystykę teoretyczną jako dział matematyki z wnioskowaniem dedukcyjnym jako narzędziem i praktykę statystyki, stosowanie metod statystycznych do rozwiązywania konkretnych zagadnień badawczych z wnioskowaniem indukcyjnym jako narzędziem. W sformułowaniu „opinie statystyków o statystyce są zwykle skrajnie optymistyczne" brak informacji, o których statystykach mówimy. W ogólności zdanie to nie jest prawdziwe.

Poświęćmy chwilę na przyjrzenie się skutkom przyjmowanej aksjomatyki, w tym momencie niezwiązanej ze statystyką. Zakładam, że każdy student – i nie tylko student – zetknął się z geometrią Euklidesa. Cały gmach tej geometrii został zbudowany około 300 r. p.n.e. na bazie pięciu aksjomatów (nazywanych także postulatami, pewnikami):

1. Od każdego punktu można poprowadzić prostą do każdego innego punktu.
2. Odcinek można dowolnie przedłużyć do linii prostej.
3. Z dowolnego środka można opisać okrąg o dowolnym promieniu.
4. Wszystkie kąty proste są równe.
5. Jeśli prosta, przecinająca dwie inne proste, tworzy z nimi po jednej stronie kąty wewnętrzne, których suma jest mniejsza od dwóch kątów prostych, to obie te proste, przedłużone nieograniczenie, przetną się po tej stronie, gdzie leżą kąty o sumie mniejszej od dwóch kątów prostych.

Aksjomat piąty w późniejszym okresie był formułowany na wiele różnych sposobów. John Pleyfair ujął go w wyjątkowo prostej formie: „Przez punkt na zewnątrz danej linii prostej może przechodzić tylko jedna prosta równoległa do niej" (Gomez, 2012).

Łobaczewski i Bolyai, niezależnie od siebie, w XIX w. zmienili aksjomat piąty: „przez punkt *P* leżący poza daną prostą *l* przechodzi więcej niż jedna prosta równoległa do prostej *l*". I zestaw aksjomatów Euklidesa ze zmienionym aksjomatem piątym stanowi podstawę geometrii hiperbolicznej. Powstała nowa geometria.

Riemann, także w wieku XIX, napisał: „Euklides twierdzi, że przez punkt poza prostą przechodzi dokładnie jedna prosta równoległa do danej prostej; Łobaczewski uważa, że takich prostych jest nieskończenie wiele; ja natomiast jestem przekonany, że takich prostych nie ma wcale". I ten zbiór aksjomatów jest podstawą jeszcze innej geometrii, geometrii eliptycznej (na sferze).

Wyniki badań Łobaczewskiego i Riemanna zostały wykorzystane przez Einsteina w definicji czasoprzestrzeni i teorii względności (Gomez, 2012).

Każda z tych geometrii jest tak samo sensowna, każda z nich jest prawdziwa, z tym że każda znajduje zastosowanie w innej sytuacji. Można powiedzieć, że w innej „rzeczywistości".

Na gruncie teorii prawdopodobieństwa i statystyki też mamy do czynienia z aksjomatyką pewnych pojęć. Na przykład pojęcie prawdopodobieństwa, podstawowe we wnioskowaniu statystycznym, zbudowane jest na bazie trzech aksjomatów, sformułowanych w 1933 r. przez Kołmogorowa. Ta aksjomatyka wyparła wcześniejsze, jako że jest bardziej nośna, lepiej nadaje się do budowy całego gmachu teorii probabilistycznej, której elementy znajdują zastosowanie w statystyce teoretycznej. Sformułowanie tych aksjomatów podaję za Fiszem (1969):

Pewnik 1. Każdemu zdarzeniu losowemu A odpowiada określona liczba $P(A)$, zwana prawdopodobieństwem zdarzenia A, spełniająca nierówność:

$$0 \leq P(A) \leq 1 \tag{0.1}$$

Pewnik 2. Prawdopodobieństwo zdarzenia pewnego równa się jedności:

$$P(\Omega) = 1 \tag{0.2}$$

Przez Ω oznaczamy przestrzeń zdarzeń elementarnych, czyli zbiór wszystkich możliwych wyników doświadczenia losowego.

Pewnik 3. Prawdopodobieństwo alternatywy skończonej lub przeliczalnej ilości zdarzeń losowych parami wyłączających się (rozłącznych) równa się sumie prawdopodobieństw tych zdarzeń.

Rozziew między teorią statystyki i zastosowaniami statystyki wyraźnie widać podczas lektury podręczników „do statystyki". O ile podręczniki pisane przez statystyków prezentują konkretne, możliwe do udowodnienia, zagadnienia – niestety, formułowane w języku matematyki – o tyle podręczniki pisane np. przez

psychologów, pedagogów i ogólniej badaczy w naukach społecznych bardzo często w ogóle nie liczą się z uwarunkowaniami teoretycznymi. Przykładowo, nieważne, jak duże prawdopodobieństwo uzyskano w teście, gdyż i tak będzie szacowana „wielkość efektu" i wykorzysta się taki wynik. Podejście takie uważam za nadmierny pragmatyzm, niekiedy za fałszowanie rezultatów badania. Opatrzenie cudzysłowem „wielkości efektu" jest wyrazem niepokoju, jaki wywołuje we mnie ten konstrukt. Będę do niego wielokrotnie wracał.

Poniżej zamieszczam konkretny przykład wnioskowania dedukcyjnego (teoria statystyki) – twierdzenie dotyczące statystyki *t*-Studenta wykorzystywanej przy porównywaniu dwóch wartości oczekiwanych (test *t*-Studenta dla prób niezależnych).

„Jeśli $\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_k$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym z wartością oczekiwaną μ i odchyleniem standardowym σ, to statystyka:

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{k+n}{kn} \cdot \frac{ns_1^2 + ks_2^2}{k+n-2}}} \tag{0.3}$$

gdzie:

$$ns_1^2 = \sum_{i=1}^{n} \left(\xi_i - \bar{x}\right)^2, \ \bar{x} = \frac{1}{n}\left(\xi_1 + \cdots + \xi_n\right) \tag{0.4}$$

$$ks_2^2 = \sum_{j=1}^{k} (\eta_j - \bar{y})^2, \ \bar{y} = \frac{1}{k}(\eta_1 + \cdots + \eta_k) \tag{0.5}$$

ma rozkład *t*-Studenta z $k + n - 2$ stopniami swobody" (Zubrzycki, 1970).

Przytoczone twierdzenie daje się udowodnić, ale czy znajduje ono zastosowanie w praktyce wykorzystywania metod statystycznych? Jest stosunkowo trudno w bezpośredni sposób wykorzystywać zarówno to, jak i inne twierdzenia statystyki teoretycznej, gdyż w praktyce niezwykle rzadko materiał empiryczny spełnia założenia twierdzenia. A jeśli założenia nie są spełnione, to twierdzenie przestaje być prawdziwym.

W tej publikacji chciałbym przedstawić Czytelnikowi pewne aspekty wnioskowania statystycznego, aksjomatyczne podstawy tego wnioskowania oraz będące konsekwencją różnej aksjomatyki problemy pojawiające się podczas stosowania metod statystycznych w praktyce. Innego typu konsekwencją problemów wnioskowania statystycznego, wynikających z różnych aksjomatyk, jest negowanie przydatności testowania hipotez przez wielu badaczy i poszukiwanie innych rozwiązań, np. ocenianie wielkości efektu, które też nie jest odbierane przez wszystkich jednoznacznie pozytywnie. Te zagadnienia również spróbuję naświetlić w jednym z rozdziałów.

Oczywiście, najpierw należy sprecyzować pojęcie wnioskowania statystycznego, co już samo w sobie jest zadaniem dość skomplikowanym. W tym obszarze

funkcjonują dwa przenikające się pojęcia. Istnieje pojęcie wnioskowania statystycznego (*statistical inference*) i pojęcie dowodu statystycznego (*statistical evidence*).

Pojęcia wnioskowania statystycznego zaczęto używać niedługo po sformułowaniu aksjomatyk testowania hipotez statystycznych, zatem mówiąc o wnioskowaniu statystycznym, będziemy traktowali je jako podejmowanie decyzji wykorzystujących metody testowania hipotez statystycznych i estymacji, zarówno punktowej, jak i przedziałowej.

Inna ważna kwestia związana z poprawnością wnioskowania statystycznego to zagadnienie mocy zastosowanego testu. Praktycznie wszystkie wykorzystywane przez nas testy statystyczne nie kontrolują prawdopodobieństwa błędu drugiego rodzaju, co znakomicie utrudnia nam poprawne wnioskowanie na podstawie wyników tych testów. Zagadnienie mocy testu i szacowania mocy testu na podstawie próby również będzie przedmiotem jednego z rozdziałów. Oczywiście, moc testu jest nierozerwalnie związana z wielkością próby, więc nie uciekniemy od kwestii szacowania wielkości próby w konkretnych analizach.

Zagadnienia poruszane w tej książce w żadnej mierze nie wyczerpują wszystkich aspektów wnioskowania statystycznego. Moim celem było zwrócenie uwagi na pewne problemy związane ze stosowaniem metod statystycznych i w konsekwencji – skłonienie badacza do refleksji nad wnioskami sformułowanymi na podstawie przeprowadzonego badania i analizy statystycznej uzyskanych danych. Chciałbym ograniczyć do minimum sytuacje, w których koronnym argumentem badacza jest: „tak wyszło z komputera" albo „tak wyszło z obliczeń". Bez względu na to, co „wyjdzie" z obliczeń, nic nie zwalnia badacza z konieczności, obowiązku – podkreślam – konieczności analizy uzyskanych wyników w terminach merytorycznych.

Praktycznie wszystkie przykłady są moimi obliczeniami w odpowiednich pakietach statystycznych. Wykorzystywane w nich były wyniki dwóch badań: wykonanych przez Bohdana Dudka, który oceniał wpływ stresu związanego z pracą na stan zdrowia pracowników służb mundurowych (Dudek, 2007) oraz przeprowadzonych przez Agnieszkę Kubot, w którym porównywano efektywność trzech rodzajów terapii w leczeniu „łokcia tenisisty" (Kubot, 2017). W przykładach wykorzystujących inne zbiory danych odpowiednie informacje podawałem w tych przykładach.

Zamieszczane w moim tekście tłumaczenia angielskojęzycznych cytatów są stosunkowo dosłowne i najczęściej zawierają nieuzasadnione (w oryginale, a więc i w tłumaczeniu) „skróty myślowe", by jak najdokładniej oddać zawarte w nich nieprawidłowości. Dlatego też, dla uniknięcia zarzutu złośliwości przy tłumaczeniu, w przypisach podaję brzmienie oryginalne.

Jeszcze drobna uwaga redakcyjna. W przytaczanych dalej przykładach będą pojawiały się oryginalne wydruki z pakietów statystycznych: SPSS 24, STATA 13 i SYSTAT 13. Starałem się możliwie niewiele ingerować w zawartość tych wydruków, choć często zawierają one zbyt dużo informacji. Zależy mi, aby Czytelnik miał możliwie pełny obraz prezentowanego zagadnienia.

Rozdział 1. Wnioskowanie statystyczne

1.1. Wprowadzenie

Opracowując wyniki badań z wykorzystaniem statystycznych metod analizy, podejmujemy decyzje będące konsekwencjami zastosowanych metod pomiaru czy klasyfikacji. Decyzje te dotyczą zarówno wyników testowania hipotez statystycznych, jak i zagadnień estymacji, głównie parametrów. Jednak kontrowersje, które budzą podejmowane decyzje odnoszą się w zasadzie wyłącznie do zagadnień testowania hipotez statystycznych. Kontrowersje te sięgają znacznie głębiej niż tylko decyzje podejmowane w wyniku testowania, gdyż wielu użytkowników metod statystycznych stawia pod znakiem zapytania także podstawy, czyli samą aksjomatykę teorii testowania hipotez statystycznych. Lambdin (2012) sugeruje w tytule swojego artykułu, że testowanie hipotez to „czarnoksięstwo". Opinię swoją opiera na fakcie ulegania kultowi istotności statystycznej przez badaczy w naukach społecznych. Niestety, w podsumowaniu artykułu brakuje propozycji innego podejścia do wnioskowania statystycznego, a jak sam autor przyznaje, jest to jedynie „wołanie o reformę" (*a call for reform*). Armstrong (2007) uważa, że testy istotności niszczą postęp w prognozowaniu. Twierdzi on, że statystyczne testy istotności są szkodliwe dla rozwoju wiedzy naukowej, ponieważ odwracają uwagę badacza od użycia **odpowiednich metod**[1]. Z kolei Loftus (1996) twierdzi, że psychologia byłaby znacznie lepszą nauką, gdyby zmienić sposób analizy danych. Zmiana sposobu analizy miałaby, według autora, polegać np. na prezentowaniu danych w postaci wykresów zamiast w tabelach z odpowiednimi prawdopodobieństwami, podawaniu przedziałów ufności (co nie jest żadnym odkryciem, chociaż nie dla wszystkich parametrów ma sens), realizowaniu metaanaliz i używanie wielkości efektu. Warto tu zwrócić uwagę, że metaanaliza jest czymś innym niż analiza danych konkretnego badania – jest to analiza opublikowanych wyników wielu badań dotyczących tego samego zagadnienia.

1 „[…] test of statistical significance are harmful to the development of scientific knowledge because they distract the researcher from the use of **proper methods**" (pogr. moje – W.Sz.).

Istnieją także opinie przeciwne, np. Häggström (2012) próbuje wyjaśnić, „dlaczego nauki empiryczne tak desperacko potrzebują statystyki?". Autor sądzi, że stosowanie metod statystycznych w dyscyplinach empirycznych jest nieuniknione i często bywa nadużywane. W konsekwencji dyskusja metodologiczna w tych dyscyplinach będzie często pociągała za sobą kwestie statystyczne, ale takie dyskusje bez udziału statystyków niosą ryzyko, że będą one po prostu nieprofesjonalne, nie będą dostarczały potrzebnych informacji. Z drugiej strony, Häggström przestrzega również przed uleganiem „kultowi istotności statystycznej". „Kult istotności statystycznej" rozumiany jest jako uznanie za ważniejsze, iż prawdopodobieństwo w teście jest mniejsze od przyjmowanego poziomu istotności, niż ewentualna ocena merytoryczna otrzymanej zależności.

Z kolei zagadnienia estymacji nie budzą raczej żadnych emocji i bardzo często opracowane wyniki analizy wyglądają tak, jakby zagadnienie estymacji nie istniało, co jest oczywistą nieprawdą. Na przykład, wykorzystując oszacowane wartości współczynników regresji, budujemy odpowiednią funkcję opisującą zależność między badanymi zmiennymi.

Wśród badaczy stosujących podczas opracowywania wyników swoich badań metody testowania hipotez statystycznych stosunkowo często można spotkać następującą opinię: im mniejsze prawdopodobieństwo w teście, tym istotniejszy wynik (silniejsza zależność). Przykładowo, Chmura-Kraemer i Kupfer (2006) piszą: „Typowy sposób przedstawiania wyników testowania hipotez statystycznych określa rezultat jako »statystycznie istotny« przy $p < 0{,}05$, co oznacza, że dane wskazują, iż dzieje się coś nielosowego. Gdy $p < 0{,}01$ dowody są bardziej przekonujące, a $p < 10^{-6}$ naprawdę bardzo przekonujące. Jednakże, chociaż wartość p pozwala określić, jak przekonująco dane świadczą przeciwko hipotezie zerowej o losowości, konkluzja zawsze przybiera formę: »zdarzyło się coś nielosowego«"[2]. Na ile nieprawdziwe jest to stwierdzenie i z czego ono wynika?

Problem polega głównie na tym, że wszystkie stosowane w praktyce testy statystyczne są tzw. testami istotności, tj. testami, które nie kontrolują prawdopodobieństwa błędu drugiego rodzaju. Wszystkie one kontrolują prawdopodobieństwo błędu pierwszego rodzaju, lecz nie kontrolując prawdopodobieństwa błędu drugiego rodzaju, uniemożliwiają nam podjęcie decyzji o przyjęciu hipotezy zerowej. Jeśli prawdopodobieństwo w teście jest większe od przyjętego poziomu istotności (najczęściej jest to wartość $\alpha = 0{,}05$), stwierdzamy, że nie ma podstaw do odrzucenia hipotezy zerowej. Praktycznie jesteśmy wówczas w sytuacji pełnej niewiedzy. Nieco lepiej, choć też nie w sposób doskonały, wygląda sytuacja, gdy prawdopodobieństwo w teście jest mniejsze od przyjmowanego poziomu istotności.

2 „As statistical hypothesis testing is typically performed, a 'statistically significant' result with $p < .05$ means that the data indicate that something nonrandom is going on. When $p < .01$, the evidence is more convincing, and $p < 10^{-6}$ very convincing indeed. However, the p value is a comment on how convincing the data are against the null hypothesis of randomness; the conclusion is always 'something nonrandom is going on' ".

Podejmujemy wówczas decyzję o odrzuceniu hipotezy zerowej (traktujemy ją jako fałszywą) i przyjęciu hipotezy alternatywnej (uznajemy ją za prawdziwą). Ale i w tym przypadku również nie mamy komfortowej sytuacji. Uznajemy, że relacje czy zależności opisane hipotezą alternatywną są prawdziwe, lecz badacza zazwyczaj zaczyna wówczas interesować, jak silne są to relacje (słowo „wpływ" rezerwuję dla relacji przyczynowo-skutkowych, a nie statystycznych).

Interpretację wielkości prawdopodobieństwa uzyskanego w teście według Chmury-Kreaemer i Kupfera (2006) można uczynić bardziej intuicyjną. Mianowicie, prawdopodobieństwo mniejsze od 0,05 będzie oznaczało, że w jednym doświadczeniu zaszło mało prawdopodobne zdarzenie, co podaje w wątpliwość prawdziwość hipotezy zerowej. Prawdopodobieństwo rzędu np. 10^{-6} oznacza, iż w pojedynczym doświadczeniu zaszło zdarzenie „prawie niemożliwe", co tym bardziej świadczy przeciwko hipotezie zerowej.

Dość powszechna interpretacja, że im mniejsze prawdopodobieństwo uzyskane w teście, tym silniejsza zależność (tutaj w terminach merytorycznych) nie ma żadnego uzasadnienia statystycznego, choć badaczy ciągle dręczy pytanie: „a jak silna jest to zależność?". Pytanie to można potraktować jako szczególną wersję ogólniejszego problemu: czy wnioskowanie statystyczne (*statistical inference*) i wnioskowanie naukowe (*scientific inference*) są tym samym? Zagadnienie to ciągle jeszcze nie zostało rozwiązane i jest przyczyną dyskusji między statystykami i badaczami stosującymi statystykę. Osobiście skłaniałbym się do opinii, że są to dwa różne, choć powiązane ze sobą, sposoby wnioskowania. Nieco więcej informacji, które uzasadniałyby moją opinię znajdzie Czytelnik w podrozdziale 4.4 (*Merytoryczne znaczenie obserwowanych różnic i wielkość efektu*).

W następnych podrozdziałach spróbuję wskazać przyczyny obecnych problemów z interpretacją wyników testowania hipotez statystycznych oraz rzeczywiste niedoskonałości istniejących rozwiązań. Informacje zawarte w bieżącym rozdziale pozwolą Czytelnikowi uświadomić sobie, dlaczego pojawiło się coś takiego, jak pojęcie wielkości efektu, natomiast pominięcie tego rozdziału nie przeszkodzi mu w wykorzystywaniu mierników wielkości efektu.

1.2. Wnioskowanie statystyczne (*statistical inference*) i dowód statystyczny (*statistical evidence, statistical proof*)

W piśmiennictwie, szczególnie angielskojęzycznym, możemy spotkać się z dwoma pojęciami opisującymi efekt statystycznej analizy danych – są to wnioskowanie statystyczne i dowód statystyczny. I tu pojawia się pytanie, jak rozumieć pojęcie wnioskowania statystycznego, a jak dowodu statystycznego? Czy są to różne pojęcia? Jeśli tak, to czym się różnią? Jeśli nie, to po co istnieją obydwa?

Należy zwrócić uwagę, że mówimy o wnioskowaniu statystycznym, czyli wnioskowaniu wykorzystującym metody statystyczne. Zatem wnioski, na tym etapie rozważań, są formułowane w terminach statystycznych. Young i Smith (2005) proponują następującą definicję wnioskowania statystycznego: „we wnioskowaniu statystycznym dane pochodzące z eksperymentu albo badania obserwacyjnego są modelowane jako obserwowane wartości zmiennych losowych, by dostarczyć pewnych ram umożliwiających wyciąganie indukcyjnych wniosków o mechanizmach działających w populacyjnych danych".

Proponowałbym pojęcie nieco bardziej intuicyjne: przez wnioskowanie statystyczne będziemy rozumieli postępowanie wykorzystujące metody statystyczne, umożliwiające uogólnienie zależności zaobserwowanych w próbie na populację generalną, z której ta próba pochodzi.

Pojęcie próby oraz sposoby jej doboru zostały krótko omówione w podręczniku Szymczaka (2018). Więcej szczegółów Czytelnik znajdzie w monografii Levy'ego i Lemeshowa (1991).

Tak rozumiane wnioskowanie statystyczne jest na tyle ogólne, że praktycznie nie zostało przypisane do żadnego z paradygmatów statystycznych.

A co z dowodem statystycznym? Bill Thompson (2007) próbuje nakreślić pewne ramy dla pojęcia dowodu statystycznego, przyznając jednak, że nie potrafimy precyzyjnie go zdefiniować. Thompson wiąże pojęcie dowodu statystycznego z eksperymentem, zauważając przy tym, iż pojęcia eksperymentu, parametru i dowodu odgrywają centralną rolę w teorii statystyki, a mimo to ich znaczenie jest często starannie pomijane. Zatem sądzę, że pojęcie dowodu statystycznego jest innym zwrotem – służącym uniknięciu powtórzeń – na określenie wnioskowania statystycznego. I zamiast korzystać z jakiegoś automatycznego miernika siły dowodu statystycznego, badacz będzie musiał przeprowadzić merytoryczną interpretację wyników analizy statystycznej i podjąć odpowiednią decyzję, już w terminach merytorycznych.

1.3. Paradygmaty statystyki

Pojęcie paradygmatu zostało zaproponowane dopiero w 1962 r. przez Thomasa S. Kuhna (Kuhn, 2009). W czasach „przedparadygmatowych" występowało pojęcie podstaw statystyki czy podstaw wnioskowania statystycznego. Poniżej przedstawię dyskusję, jaka toczyła się na temat podstaw statystyki (*foundations of statistics*) w latach 1958–1962 po ukazaniu się książki Savage'a w 1954 r. (Savage, 1954). Ale także w późniejszym okresie zdarzały się artykuły, których autorzy, dyskutując o podstawach statystyki, nie używali pojęcia paradygmatu, np. Efron (1978) czy Freedman (1995/1996).

Autorzy analizujący podstawy statystyki na ogół zgadzają się, że podstawy statystyki są kontrowersyjne i zmienne. Jednakże są one częścią podstaw nauki w najszerszym sensie i mimo niezbędności wykorzystywania w statystyce narzędzi matematycznych, jej podstawy mogą być rozważane w aspekcie filozoficznym (Savage, 1958). Z innego punktu widzenia, statystyka wydaje się trudna również dla matematyków – być może z powodu nieosiągalności tradycyjnej metody przedstawiania wyników w matematyce poprzez twierdzenie-dowód. Słabym pocieszeniem wydaje się fakt, iż statystyka jest trudna i dla statystyków (Efron, 1978).

Toczącej się od kilkudziesięciu lat (niektórzy uważają, że od ponad dwustu lat) dyskusji na temat podstaw statystyki nie widać końca. W chwili obecnej wcale nie jesteśmy bliżsi ujednolicenia podejścia do metod statystycznych niż byliśmy kilkadziesiąt lat wcześniej. Nadal toczą się spory między wyznawcami podejścia częstotliwościowego i wyznawcami podejścia bayesowskiego. Nieco inaczej granica ta przebiega między wyznawcami obiektywności i subiektywności w statystyce. W grupie zwolenników podejścia częstotliwościowego też występują różnice między opowiadającymi się za „wnioskowaniem indukcyjnym" według teorii Fishera i „postępowaniem indukcyjnym" według teorii Neymana-Pearsona. Dyskusje te toczą się wśród matematyków i statystyków. Spróbujmy wyobrazić sobie, co dzieje się wśród badaczy wykorzystujących metody statystyczne do opracowywania wyników swoich badań, którzy nie mają wiedzy matematycznej. W rozdziałach trzecim i czwartym przedstawię pewne ich propozycje rozwiązania problemu.

Analiza statystyczna nie zajmuje się badaniem zjawisk deterministycznych, jej przedmiotem są zjawiska losowe. Aby w pewien sposób „okiełznać" nieprzewidywalność pojawiania się takich zdarzeń, niezbędna jest jakaś miara pozwalająca – z lepszym lub gorszym skutkiem – przewidywać nieprzewidywalne. Taką miarą w statystyce, przynajmniej na pierwszym etapie jej rozwoju, było prawdopodobieństwo. Kłopot z tą miarą polega jednak na tym, że nie posiadamy intuicji prawdopodobieństwa. Skutkuje to np. takimi stwierdzeniami: „Jeśli prawdopodobieństwo jakiegoś zdarzenia jest prawie równe 1, to z dużym stopniem pewności zdarzenie to pojawi się w pojedynczej próbie" (Papoulis, 1972). Papoulis tym stwierdzeniem pokazuje, na czym polega problem z prawdopodobieństwem. Bo cóż oznacza duży stopień pewności? Jest to po prostu inna nazwa stosunkowo dużego prawdopodobieństwa. Zatem cytowane zdanie nic nie wyjaśnia. I trzeba się zgodzić, że „teoria statystyczna, która jest ścisłą dyscypliną rozwiniętą z jasno sformułowanych aksjomatów, jest powiązana ze zjawiskami fizycznymi tylko poprzez nieścisłe terminy" (Papoulis, 1972). Czy jednak należy zgodzić się ze stwierdzeniem, że statystyka jest dyscypliną rozwiniętą z jasno sformułowanych aksjomatów? Raczej różne statystyki są rozwijane z jasno formułowanych różnych zbiorów aksjomatów.

Ale wróćmy do zagadnień prawdopodobieństwa zdarzenia. Brak intuicji prawdopodobieństwa zdarzenia spowodował powstanie kilku definicji prawdopodobieństwa, co doskonale utrudnia późniejsze wykorzystywanie tego pojęcia w analizach statystycznych. Możemy wyróżnić:

- definicję aksjomatyczną (Kołmogorow),
- definicję klasyczną (Laplace),
- definicję wykorzystującą częstości względne (von Mises),
- prawdopodobieństwo jako miarę przekonania.

Definicja aksjomatyczna (Kołmogorow, 1933)

Każdemu zdarzeniu (zdarzeniu losowemu) A przyporządkowana jest liczba $P(A)$, spełniająca następujące warunki:

1) $P(A)$ jest nieujemna; $P(A) \geq 0$,
2) prawdopodobieństwo zdarzenia pewnego jest równe jedności; $P(\Omega) = 1$,
3) prawdopodobieństwo alternatywy (sumy mnogościowej) skończonej lub przeliczalnej ilości zdarzeń losowych parami wyłączających się jest równe sumie prawdopodobieństw tych zdarzeń:

$$P(\bigcup_k A_k) = \sum_k P(A_k);\ \ A_i \cap A_j = \emptyset\ \ i,j = 1,2,\cdots,k;\ i \neq j \tag{1.1}$$

Wzór ten można zapisać w nieco innej postaci:

$$\begin{aligned} &P(\mathrm{A}_1 \cup A_2 \cup \cdots \cup A_k \cup \cdots) = \\ &= P(A_1) + P(A_2) + \cdots + P(A_k) + \cdots;\ A_i \cap A_j = \emptyset\ \ i,j = 1,\cdots,k;\ i \neq j \end{aligned} \tag{1.2}$$

Oprócz własności prawdopodobieństwa wynikających bezpośrednio z aksjomatycznej definicji, czyli własności, iż prawdopodobieństwo zdarzenia pewnego jest równe jedności:

$$P(\Omega) = 1 \tag{1.3}$$

oraz że prawdopodobieństwo alternatywy (sumy mnogościowej) skończonej lub przeliczalnej ilości zdarzeń losowych parami wyłączających się jest równe sumie prawdopodobieństw tych zdarzeń, warto dodać jeszcze jedną: prawdopodobieństwo zdarzenia niemożliwego jest równe zero:

$$P(\emptyset) = 0. \tag{1.4}$$

Tak zdefiniowane prawdopodobieństwo w żaden sposób nie poprawia intuicji tego pojęcia. Jest wygodne, eleganckie i efektywne dla rozwijanej na jego podstawie teorii probabilistycznej, lecz nie ułatwia (a nawet nie umożliwia) interpretacji podczas oceny rezultatów analiz statystycznych.

Klasyczna definicja prawdopodobieństwa (Laplace, 1812)

Klasyczna definicja prawdopodobieństwa sformułowana przez Laplace'a znajduje zastosowanie tylko w przypadku skończonych zbiorów zdarzeń elementarnych.